LEVEL II

ARI TECHNICAL REPORT
TR-78-A31

Criterion-Referenced Measurement in the Army: Development of a Research-Based, Practical, Test Construction Manual

by

Richard B. Pearlstein and Robert W. Swezey

APPLIED SCIENCE ASSOCIATES, INC.
Box 158
Valencia, Pennsylvania 16059

SEPTEMBER 1978

Contract DAHC 19-74-C-0018

Contracting Officer's Technical Representative Angelo Mirabella
Unit Training & Evaluation Systems Technical Area, ARI

Prepared for

U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
5001 Eisenhower Avenue
Alexandria, Virginia 22333

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

JOSEPH ZEIDNER
Technical Director

WILLIAM L. HAUSER
Colonel, US Army
Commander

Research accomplished under contract
to the Department of the Army

Applied Science Associates, Inc.

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

b

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Tk-78-A31
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): CRITERION-REFERENCED MEASUREMENT IN THE ARMY: DEVELOPMENT OF A RESEARCH-BASED, PRACTICAL, TEST CONSTRUCTION MANUAL
5. TYPE OF REPORT & PERIOD COVERED: Final Report, 17 Dec 1973 to 16 Dec 1974
6. PERFORMING ORG. REPORT NUMBER: 308-AR18(2)-FR-1174-RBP
7. AUTHOR(s): Richard B. Pearlstein; Robert W. Swezey
8. CONTRACT OR GRANT NUMBER(s): DAHC19-74-C-0018
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Applied Science Associates, Inc., Box 158, Valencia, PA 16059
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2Q763743A773
11. CONTROLLING OFFICE NAME AND ADDRESS: U.S. Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333
12. REPORT DATE: September 1978
13. NUMBER OF PAGES: 64
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): -- ASA-308-ar18(2)-fr-1174-rbp
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited. ARI Tk-78-A31
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report): --
18. SUPPLEMENTARY NOTES: Technically monitored by Angelo Mirabella, Unit Training and Evaluation Systems Technical Area, Army Research Institute. See also "Guidebook for Developing Criterion-Referenced Tests," AD A014 987.
19. KEY WORDS (Continue on reverse side if necessary and identify by block number):

Criterion-Referenced Tests (CRT)	Testing
Performance Tests	Performance
Objectives	Validity
Training	Reliability

20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

This final report summarizes activities conducted to develop a Criterion-Referenced Tests (CRTs) Construction Manual. Major accomplishments were the preparation of a written review of the literature on Criterion-Referenced Testing, identification of needed research to help achieve a more consistent, unified Criterion-Referenced Test (CRT) Model, and development of an easy-to-use, 'How-to-do-it' manual to assist Army test developers in the construction of CRTs.

(continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

032 170

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. (continued)

In order to accomplish these objectives, the project encompassed the following activities:

1. A survey of the literature on Criterion-Referenced Testing conducted in order to provide an information base for development of the CRT Construction Manual.

2. Visits to selected Army Posts to review the present status of Criterion-Referenced Test construction and application in the Army. Interviews conducted during these visits provided information which aided in making the CRT Construction Manual practicable and usable, and in identifying problems with Criterion-Referenced Testing that require further research.

3. Preparation of an interim report based upon the first two tests and upon review by experts in the Criterion-Referenced Testing field.

4. Preparation of a draft CRT Construction Manual.

5. Revision of the draft manual, based upon feedback from expert reviews.

6. Conduction of a field review of the revised manual, in which selected Army personnel used the revised manual to construct CRTs. These personnel completed evaluation during the field review indicating the utility of the manual and problems encountered with its use. In addition, other Army personnel functioning in supervisory capacities also reviewed the manual.

7. Final revision of the CRT Construction Manual, based upon the supervising of the field review.

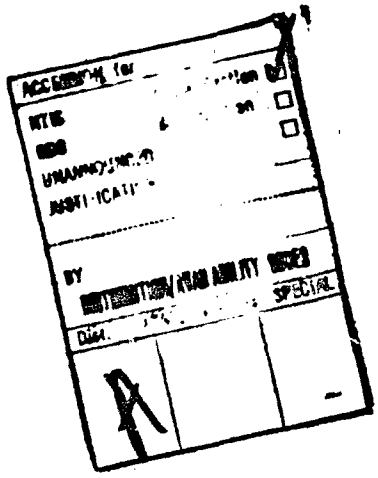


TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF FIGURES

Acknowledgments

This report has been prepared by Applied Science Associates (ASA) under Contract No. DAHC19-74-C-0018 with the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). The Principal Investigator was Dr. R. W. Swezey. Dr. R. B. Pearlstein was Lead Scientist on the tasks performed under this contract.

The authors wish to mention contributions made by a number of individuals. Dr. Angelo Mirabella, Work Unit Leader in the Unit Training Standards and Evaluation Unit of ARI, and Contract Officer's Technical Representative on this project, provided invaluable inputs and feedback during the conduct of this project. Mr. Gary Boycan and Dr. Rick Steinheiser, both members of the COTR's Work Unit, were extremely helpful and informative. Mr. Eugene Johnson and Mr. Ken Epstein of that Unit also made helpful inputs.

Major Edgar D. Maddox and Captain Frank Dowling of the U.S. Army Combat Arms Training Board have aided greatly in arranging liaison between the contractor and various Army personnel. They also provided useful feedback on several draft products resulting from this project.

Miss Alesia W. Getchell and Dr. William Ton, both of ASA, provided helpful literature review and data analysis services. Assistance in data collection activities was furnished by Mr. T. K. Elliott, Dr. A. P. Chenzoff, and Dr. A. L. Pinkus of ASA.

Dr. William Freeman at Fort Benning, Georgia; Mr. James Klaes at Fort Bliss, Texas; Dr. Charles Jackson at Fort Knox, Kentucky; Mr. Curtis McBride at Fort Sill, Oklahoma; Captain Jerry Ford at Ford Ord, California; and Mr. Robert Thompson at Fort Gordon, Georgia assisted in scheduling on-post data collection activities, and also supplied much helpful commentary and advice at several stages of the project.

Mr. James L. Sherrill of Fort Benjamin Harrison, Indiana provided a useful and constructive critique of the draft manual. Dr. Harold Edgerton of San Diego, California provided highly constructive reviews of both the survey of literature on criterion-referenced testing, and of the draft criterion-referenced test construction manual. Dr. George Macready of the University of Maryland, and Dr. Robert Branson of Florida State University furnished useful review comments on the draft CRT Construction Manual. Dr. Robert Glaser of the University of Pittsburgh has also made valuable contributions to this project.

Introduction

This final report summarizes activities conducted under a contract to develop a Criterion-Referenced Test (CRT) Construction Manual. Major objectives accomplished by the project were the preparation of a written review of the literature on Criterion-Referenced Testing, identification of needed research to help achieve a more consistent, unified criterion-referenced test model, and development of an easy-to-use, "how-to-do-it" manual to assist Army test developers in the construction of CRTs.

In order to accomplish these objectives, the project encompassed the following activities:

1. A survey of the literature on criterion-referenced testing conducted in order to provide an information base for development of the CRT Construction Manual.

2. Visits to selected Army posts to review the present status of criterion-referenced test construction and application in the Army. Interviews conducted during these visits provided information which aided in making the CRT Construction Manual practicable and useful, and in identifying problems with criterion-referenced testing that require further research.

3. Preparation of an interim report based upon the first two tasks and upon review by experts in the criterion-referenced testing field.

4. Preparation of a draft CRT Construction Manual.

5. Revision of the draft manual, based upon feedback from expert reviews.

6. Conduct of a field review of the revised manual, in which selected Army personnel used the revised manual to construct CRTs. These personnel completed evaluation packages during the field review indicating the utility of the manual and problems encountered with its use. In addition, other Army personnel functioning in supervisory capacities, also reviewed the manual.

7. Final revision of the CRT Construction Manual, based upon the findings of the field review.

This report fulfills the contract requirements for a final report summarizing project activities.

Executive Summary

Part 1 of this report describes procedures used for reviewing the technical and theoretical literature in the areas of criterion-referenced testing. Sources of the literature reviewed, search strategies, and topics covered are described.

Part 2 summarizes positions on theoretical and technical aspects of CRT construction and use, based upon the state-of-the-art of criterion-referenced testing as reflected in the literature review. These positions were used as the bases for the procedures presented in the CRT Construction Manual developed during this project.

Part 3 is a brief summary of the methodology used to survey the application of criterion-referenced testing techniques in the Army. Information was collected to supplement the literature search and review, to provide detailed material on current CRT development and use in the Army, and to obtain information concerning attitudes on, and opinions about, criterion-referenced measurement, held by Army testing personnel. Listed in this section are topics covered by the survey. Development of the Interview Protocol used in the survey is also described, along with a quick overview of the various types of personnel who participated in the survey.

Part 4 presents a summary and discussion of the results from the field survey of CRT development and use in the Army. General patterns in test construction processes which became apparent during the survey are discussed. Results of the survey are indicated through an analysis of quantitative data collected during interviews, and through a discussion of qualitative comments, opinions, and anecdotal information recorded during the interviews. Problems observed in the development and use of CRTs by the survey teams are described, and areas where changes may prove beneficial to the Army are mentioned.

Part 5 describes the development of the CRT Construction Manual. Objectives on which the manual is based are listed, and review and revision procedures are discussed.

Part 6 describes the way in which the revised draft CRT Construction Manual was evaluated in the field, and the results of the field evaluation. Additionally, this section presents a discussion of the field evaluation findings, in terms of implications for further refinement of the manual, Developing Criterion-Referenced Tests.

Part 7 presents recommendations for future research on, and implementation of, criterion-referenced measurement in Army applications.

Appendix A presents the final version of the Interview Protocol used in the Army CRT survey, while Appendices B and C are reproductions of materials used in the field evaluation of the CRT Construction Manual. Appendix D consists of tallies of the data received from the field evaluation, and median response values.

Part 1

Procedure for Reviewing the Literature on Criterion-Referenced Testing

During conduct of this project, ASA reviewed the technical and theoretical literature on criterion-referenced testing. The starting point for this literature search was a data base, developed by ASA, consisting of approximately 2,700 abstracts and evaluations of journal articles, technical reports, military training literature, and books on instructional system development, including criterion-referenced testing. During the development of this data base, nearly 12,000 documents were reviewed, and the most relevant were abstracted and evaluated. Journals reviewed included the American Educational Research Technology, Journal of Educational Research, Journal of Programmed Instruction, Psychological Record, and many others, most of which were searched as far back as 1952.

The data base additionally included sources identified by several computer searches, including an ERIC search, two DDC searches, a packaged MEDLARS search, and a search of the HumRRO KWOC Index. All searches used keywords such that references pertinent to criterion-referenced testing were likely to have been captured.

ASA used this data base as follows:

1. The data base was reviewed to select all references directly relevant to criterion-referenced testing.

2. References contained in the literature selected as being directly relevant were followed-up, thereby expanding the data base documents concerning criterion-referenced testing.

3. Additional educational literature not covered adequately during the creation of the original, instructional system development data base, was reviewed, and appropriate documents were added to the criterion-referenced testing data base.

4. All documents in the criterion-referenced testing data base were reviewed, and important points on methodology, results, and critiques were documented in a cross-referenced index file.

5. A review of the literature was prepared, based on the cross-referenced index.

view of literature was submitted to subject matter experts, including Dr. Robert Glaser of the University of Pittsburgh and Dr. Harold Edgerton of San Diego, California. These experts identified gaps in the review, and suggested additional sources of information on criterion-referenced testing, which were subsequently reviewed.

The modified review of literature was submitted to ARI. Following feedback from the COTR, the literature search was expanded, and further information was transmitted to ARI.

A review of the literature on criterion-referenced testing included the following topics:

- Reliability and Validity
- Construction Methodology
- Fidelity
- Issues Related to CRT Construction
- Mastery Learning
- Establishing and Classifying Instructional Objectives
- Developing Test Materials and Item Sampling
- Quality Assurance
- Designing for Evaluation and Diagnosis
- Establishing Passing Scores
- Uses of CRT in Non-Military Education Systems
- Military Uses
- Indirect Approach to Criterion-Referencing
- Using NRTs to Derive CRT Data
- Considerations for a CRT Implementation Model
- Cost-Benefits Considerations

The literature review itself appears in the Interim Report prepared under Contract No. DAHC19-74-C-0018 with the U.S. Army Research Institute for the Behavioral and Social Sciences, entitled Criterion-Referenced Testing: A Discussion of Theory and of Practice in the Army and is authored by Robert W. Swezey, Richard B. Pearlstein, and William H. Ton.

The following section of this final report summarizes conclusions concerning the state-of-the-art of criterion-referenced testing, based upon information contained in the literature review.

Part 2

Brief Summary of the State-of-the-Art in Criterion-Referenced Testing

The purpose of this section is to describe positions on theoretical and technical aspects of CRT construction and use, based upon the state-of-the-art of CR testing as reflected in the ASA literature review (Swezey, Pearlstein, and Ton, 1974). These positions were used as the bases for the procedures presented in the CRT Construction Manual. Positions are presented sequentially for the following topics:

1. Design considerations and CRT use
2. Construction methodology and related issues
3. CRT administration and scoring
4. Reliability and validity.

Design Considerations and CRT Use

Among the major considerations in CRT construction is the way in which specific uses may affect test design. Test design may vary in several related fundamental respects, such as the basis upon which test items are constructed and selected. In CR testing, items are generally developed from an analysis of tasks to be performed and from attempts to operationally define the behaviors required. This is not necessarily the case in norm referenced (NR) testing. The manner in which scores are interpreted and used also differentiates CRTs from NRTs. In CR testing, scores attained by examinees are interpreted against an external, absolute standard--as opposed to the distribution of scores attained by other examinees; which is the case with NRTs.

It must first be decided whether a CRT, as opposed to a NRT, is appropriate. CRT scores do not lend themselves to ordering individuals along a continuum, thus if the primary use of test results is to select among individuals for promotion, special honors, etc., CR testing is contraindicated. Whenever information is desired for purposes of comparing examinees, NR testing appears to be more appropriate than CR testing. This applies to tests of achievement, knowledge, and performance.

CR testing is usually the technique of choice when evaluations are to be made on the basis of an individual's achievement of specific objectives. Here the primary question of interest is: "How well can an individual perform relative to an external standard?", rather than: "How well does an individual do compared to others?".

Cost Effectiveness

CRTs may be more expensive to develop and administer than NRTs, in terms of absolute costs. CRT-specific development costs are due largely to the need for carefully deriving and specifying objectives, while additional administration costs may result from the necessity of comparing examinee performance to external standards. Nevertheless, CR testing may well be more cost-effective in the long run, if there is a genuine need to ascertain an individual's ability to perform a specific task.

Indirect approaches to criterion-referencing, by correlating symbolic performance and/or job knowledge test results with performance measures, may be an approach to alleviating the high costs of CRTs. Such approaches involve the development of two tests at different levels of fidelity for each objective, and subsequent validation of the indirect measures against the performance measures. Justification for these approaches center on savings in administration time and costs.

Development of direct CRTs appears justified, desirable, and cost-effective, if there is a need to ensure that individuals will be able to perform adequately on the tasks for which they are being trained. When there is a need for ensuring minimal, absolute levels of performance, CR testing is the approach of choice.

Screening and Diagnosis

CRTs are applicable for use as screening devices in cases where there is a possibility that individuals may be able to perform tasks without training. If a person can achieve the criterion level on a CRT, he should be able to enter the job without intervening training. Similarly, CRTs may be used to determine the appropriate point in a training cycle for an individual to commence training.

CRTs may also be used as diagnostic aids. Persons achieving the criterion level might be channeled into advanced instruction, or remediation might be suggested for those falling below criterion level on certain objectives. CR testing for diagnostic purposes is likely to be more difficult and more expensive than CR testing for achievement of objectives, because detailed documentation on the examinees' behavior is required. This may necessitate more examiners and/or more elaborate schemes for collecting data.

Evaluation of Instructional Programs

Aside from the assessment of individual performance against absolute standards, CRTs may also be used to evaluate instructional programs. Here, the primary question of interest is: "Has my instructional

program taught what it is supposed to teach?". NR testing is less appropriate for such an application than is CR testing, since wide score ranges before and after administration of the instructional program are not necessarily germane to the question of interest. CRTs designed for this application are presumably based directly upon instructional objectives since the basic question is whether or not the program has successfully taught performance compatible with the instructional objectives. CRTs thus provide data having direct relevance to the question.

Construction Methodology and Related Issues

Due to the relative recency of the CR testing concept, many theoretical and practical aspects of CRT construction methodology are not so well defined as is the case for NRTs. Additional sophistication in CRT construction methodology must await further research on theoretical issues, and results from more extensive attempts at CRT implementation. Nevertheless, some general "do's and don'ts" for CRT construction can be extracted from the methodological literature.

Task Analysis

First, CRT construction requires careful analysis of the tasks comprising the test's subject. While conduct of the task analysis itself may be outside the test developer's domain, the test developer must obtain analytic data on: (1) skills and knowledges necessary for task performance, (2) required performances stated in behavioral terms, (3) criteria associated with each identified performance, and (4) conditions under which the tasks must be performed.

Without these data, the test developer cannot adequately define objectives, and consequently cannot match test items to objectives. Nor can he ensure the content validity of the test. If usable CRTs are to be constructed, task analyses are necessary prerequisites.

Preparing Objectives

Preparing objectives is one of the first formal steps in constructing a CRT. Mager (1962) has documented a useful procedure for formatting these objectives. Mager's suggestions for structuring objectives also appear appropriate. Information to be used in preparing objectives is best derived from thorough task analytic data.

If the test developer's input includes a list of unitary objectives--objectives covering separate, single tasks--as is assumed in the case of the CRT test construction process presented in the ASA Manual, the test developer's primary task is to match test items to these objectives.

The test developer must assume that objectives are properly matched to the actual job tasks. If this assumption is violated, the resulting CRT will lack content validity. If however, the assumption is accurate, and the developer properly matches items to objectives, content validity will be achieved. Thus, the test developer must be knowledgeable about appropriate formats and quality standards for objectives in order to make an adequate assessment of their suitability for CRT development.

Matching Items to Objectives

Mager (1973) has provided a sound plan for matching CRT items to objectives. Mager's plan involves matching performances and conditions stated in, or implied by objectives, with corresponding item performances and conditions. Mager's plan omits a procedure for matching standards among objectives and test items, however implies that standards should also be matched.

The test constructor's task is to create test items that are congruent with objectives. To the extent that objectives are "fuzzy," the test constructor cannot create appropriate items. It is recommended that he send fuzzy objectives back to their originator, annotating their difficulties and requesting a reconsideration.

When the test developer has received an adequate objective (or set of objectives) for which a test is to be constructed, a number of factors must be considered before items are matched to objectives. These factors include: practical constraints in the testing situation, test fidelity, test format, and number of items required to test a given objective.

Practical constraints must be systematically assessed before test items can be constructed so that the items can be built with performance indicators which are suitable for such considerations as: testing conditions, tester availability, time availability, facility and equipment availability, etc. These considerations obviously impact on test fidelity. CRT items should be constructed at the highest level of fidelity practicable, consistent with situational constraints. In cases where critical objectives are to be tested, special care must be taken to develop sufficiently high fidelity items so that critical task mastery can be accurately assessed.

Selecting Among Objectives

The tactic of selecting among objectives, that is, randomly testing a subset of objectives, may be used in some instances, as long as trainees do not know the subset to be tested. This tactic must not be used when critical objectives are involved. For objectives of a

non-critical nature, selection may be used to overcome practical constraints imposed by the testing situation, without necessitating modification of objectives. Selection among objectives should never be done when it is necessary to certify that individuals qualify on all objectives.

Number of Items

No hard and fast rules for specifying the number of items to be created for a given objective exist. It is recommended that as many items as test situation time availability will permit, within limits suggested by considerations of motivational and fatigue factors, should be included. As Graham (1974) has noted, "even for highly homogeneous tests, four or five items may be necessary to minimize classification errors." Thus, even for CRTs measuring a single, well-specified objective with few confounding factors, additional items may help to reduce measurement error. For more heterogeneous tests, the desirability of having extra items may be even more pronounced.

Format

Test format may, in many cases, be largely dictated by objectives. Certain objectives for example, may require hands-on performance testing. Such things as number of items to be included, and practical constraints such as time and manpower availability, may also help determine format--e.g., a situational item, multiple-choice format might be the only feasible way of testing some sets of objectives. A general guideline might be based on Edgerton's (1974) suggestion, that item styles not be mixed in the same test, so as to avoid measuring "test taking skill" instead of subject matter competence.

Item generation rules, such as "item forms" and "facets" are not yet sufficiently researched to warrant use by personnel who are not sophisticated in psychometrics. Hence, for objectives that may be tested by an unlimited number of items, such as those dealing with concepts, the best suggestion that can be offered testing personnel at this time, is to be sure that each item matches the objective it tests.

Item Pools

After the test developer has considered such factors as fidelity, number of items, etc., items can be matched to objectives using principles similar to those advanced by Mager (1973). The test developer should construct a pool of items considerably larger than the number required for the test, so that the best items can be selected. Items are then constructed at the level of fidelity and in the format previously determined.

Item Analysis

Traditional item analysis techniques, like other statistical techniques developed in conjunction with NR testing, have limited applicability for CR testing (due to restricted ranges of score variance in CRTs). Although recent studies have suggested techniques for increasing variance of CRT scores (e.g., Haladyna, 1973; Woodson, 1973) these techniques are "experimental," and it is not yet appropriate to apply them as a matter of course. Consequently, until additional research develops and refines new approaches to item analysis appropriate for CR testing, a simple index which relies on the use of "masters" and "non-masters" (e.g., those who are beginning training and those who have completed training) appears to be an appropriate technique.

"Masters" and "non-masters" are tested and their patterns of pass and fail on the items are recorded. ϕ coefficients are computed using four-fold tables ("master"-"nonmaster," pass-fail) for each item. Good items are those which are passed by "masters" and failed by "nonmasters." Items are poor if there is little difference on pass-fail patterns between "masters" and "nonmasters," or if more "nonmasters" than "masters" pass them. Low or negative ϕ coefficients act as warning flags. Items receiving low coefficients should either be thrown out or, at least, reconsidered carefully before inclusion in a CRT. These warning flags are relevant if the pool of items is homogeneous, or if it is composed of items testing several objectives.

All items should also be reviewed via peer evaluation, subject matter expert evaluation, and by appropriate test evaluation units. Care must be exercised to ensure that all objectives are represented by the proper number of items, as determined previously. Item balance among disparate objectives measured by the same test should be maintained as planned.

CRT Administration and Scoring

Administration

Like all tests, CRTs must be administered under standardized conditions. CRTs should include accompanying documentation which specifies: (1) test administration conditions; (2) instructions; (3) administration procedures (including how to handle questions, how to check and set up test supplies and equipment, etc.); (4) circumstances for excusing examinees from the test, due to illness, fatigue, etc.; (5) environmental circumstances under which test administration should be cancelled; and (6) scoring procedures.

Test administrators must be trained to follow specifications precisely. Since specifications will apply to any test, documentation accompanying a specific CRT need not necessarily be extremely detailed--

except for special requirements such as setting up the test facility, and test scoring.

Scoring

Test scoring procedures must be developed during the test construction process, since they will generally vary as a function of the type of CRT. There are a number of interrelated decisions that must be made concerning scoring. These include:

1. Objectivity of scoring
2. Process vs product scoring methods
3. Type of scoring (go/no-go, rating scales, etc.)
4. Cut-off points
5. Non-interference vs assist methods.

Objectivity

Every attempt should be made to maximize objectivity in scoring CRTs. In low fidelity tests, such as those using multiple-choice formats, objectivity is apparent. (Such tests can be computer-scored.) In higher fidelity CRTs, it is relatively simple to maximize objectivity for hard-skill subjects, however soft-skill areas, such as tactics, leadership, etc. are more difficult to test objectively. To the extent that objectivity is not achieved, reliability is attenuated. Efforts must be made to specify soft-skill objectives precisely, so that appropriate items (with associated objective scoring procedures) can be prepared. Even in the best of circumstances, however, soft-skill CRTs will probably have less objective scoring guides than will tests of hard-skill subjects. One way to maximize objectivity in soft-skill CR testing is to require several raters to assess each individual. Inter-rater reliability can then be calculated. If low inter-rater reliability is found consistently, the test should be revised.

Process-Product

R. G. Smith's (1965) guidelines for determining process versus product measurement appear adequate, with slight modifications. That is, product measurement is always appropriate if the objective specifies a product. When a product measure is called for, it should be incorporated into the objective, and carried over into the test items. Product measures are called for when:

(a) the product can be measured as to presence or characteristics

(b) the procedure leading to the product can vary without affecting the product.

Process measurement is indicated when the objective specifies a required sequence of performances which can be observed, and the performance is as important as the product. Process measurement is also appropriate in cases where the product cannot be measured for safety or other constraining reasons.

There may also be situations where both process and product measurement are appropriate for a given objective. Following are several examples of conditions that may call for both product and process measurement:

(a) Although the product is more important than the process(es) which lead to its completion, there are critical steps which, if misperformed, may cause damage to equipment or injury to personnel.

(b) The process and product are of similar importance, but it cannot be assumed that the product will meet criterion levels.

(c) Diagnostic information is needed. (By having process as well as product measures, information as to why the product does not meet the criterion can be obtained.)

When both process and product measures are obtained for a specific objective, scoring must follow the criterion specified by the objective. That is, if the criterion specifies only a product, then process scores should not be used to assess achievement of the criterion.

Type of Scoring

The type of scoring system employed must be appropriate for the objective. If the objective specifies an action or product, a go/no-go scoring system should be used (either the action occurs in the proper sequence or it does not; either the product results or it does not). If the objective specifies characteristics of a criterion-level product or action, a rating scale or other form of point assignment is indicated. Point assignments must be made on an explicit, well-defined basis for each item. For rating scales, inter-rater reliability must be high. Point assignments must be tied to criterion levels specified in the objective.

Cut-Off Points

Cut-off levels should reflect mastery of the objective to the extent required. Since factors other than ability to perform a task (such as careless errors, measurement errors, etc.) may affect an individual's score, cut-off levels are often set somewhat below 100 percent. If, for example, an objective calls for multiplication of two

four-digit numbers, the criterion might specify performing 10 such sets within five minutes, achieving the correct answer in at least eight cases. Thus, the cut-off score of 8 (below 8 = fail) reflects an arbitrary definition of mastery. True mastery would require 10 out of 10.

Graham (1974) has made some valuable suggestions concerning the setting of cut-off points. The cut-off, basically, should discriminate masters from non-masters. However, as item domains become more broad, more heterogeneous item sets are required. Thus, the confounding influence of skills and knowledges which are not directly related to objectives increases. For tests measuring objectives having broad domains (or several objectives with different domains) the overlap between mastery and non-mastery scores consequently widens.

When little overlap occurs between mastery and non-mastery scores (as is the case for tests measuring a single objective with a relatively restricted domain) setting a cut-off score is less critical. The cut-off point should reflect the standard specified by the objective, and can do so without falling into the zone of overlap between masters and non-masters, since this zone, by definition, is either narrow or non-existent. On the other hand, if the overlap is wide, the point at which the cut-off score is set, is critical. Wherever the cut-off score is set, there will be some misclassification. In such cases, there are two considerations. First, objectives must be specified precisely, with item domains as restricted as possible, in order to narrow the mastery-nonmastery overlap. When achievement of several objectives of disparate nature are measured by a single test, separate scores for each objective's item set should be obtained, each with its own cut-off. However, for end-of-course or end-of-cycle exams which assess high levels of skill and knowledge integration, a single cut-off may be set, since what is to be evaluated is a cluster of skills and knowledges applied in combination.

Second, costs of false positives and false negatives must be considered. If the costs for false negatives are relatively high (e.g., manpower needs are critical) the cut-off score might justifiably be lowered. If the costs of false positives are high, then cut-off scores must remain high. In any case, when performance on critical tasks is tested, cut-off points must be kept high enough to reflect the standards specified in the objectives for those tasks.

Assist vs Non-Interference

In general, a non-interference method of test administration is preferred over an assist method, in CR testing applications. In the assist method, the examinee is scored no-go for a missed item, corrected, and then allowed to proceed. A major problem here, is that if the criterion requires an examinee to complete a chain of steps, he should be tested on to his ability to do so. On the job, the examinee will have

to complete the chain of steps correctly, with no help. There are however, cases in which an assist scoring technique can be profitably used. These involve uses of CR testing for diagnosis. In such cases, the trainee is permitted to complete a chain of steps and given assistance on those which he cannot perform adequately. He is typically scored no-go for steps where he is assisted. The record of no-go steps is a useful diagnostic tool--remediation can concentrate on missed steps. Such records may also be useful for evaluating instructional material, especially if many examinees have similar patterns of no-go items.

Reliability and Validity

Reliability

Techniques for assessing CRT reliability are, for the most part, either not fully developed or are based on questionable assumptions. (For example, see Livingston, 1972; Oakland, 1972; Haladyna, 1974; and Woodson, 1974.) The need for additional work in the area of CRT reliability continues to be a pressing one.

A practical solution is to assess test-retest reliability of CRTs, a procedure which does not depend on internal consistency, and which increases the variability of test results, because of the two test administrations required. The ϕ coefficient is useful for analyzing the resulting fourfold (first administration-second administration, pass-fail) data. ϕ values less than +.50 would indicate unacceptable test-retest reliability for CRTs.

Validity

Content validation is an especially appropriate method in CRT applications. A CRT is content valid if the test items are carefully based on the performances, conditions, and standards specified in the objectives and if the test items appropriately sample objectives. (Of course, the objectives themselves must be sound.) Thus, in most instances, careful test construction will, itself, enable the development of content valid CRTs. However, in instances where low fidelity CRTs are constructed, it may be more difficult to determine content validity, since the items are not likely to be precisely matched to objectives. In such cases, there are two additional types of criterion-related validation that are well-suited to CRTs: concurrent validity and predictive validity.

In determining concurrent validity, CRT results are compared with an outside measure of the behaviors tested by the CRT. This outside measure must be the best available assessment of performance on the objective(s) in question. The assessment of concurrent validity, involves individual assessment via the CRT and the outside measure close

together in time (concurrently). ϕ again is used on the four-fold data (CRT-other measure, pass-fail)

Predictive validity involves the same assumptions. The outside measure must be an accurate measure of the performance in question, or the validation will be meaningless. Predictive validity is calculated the same way, except the outside measure is taken at a later time--i.e., when the individuals are actually performing the job for which they've been trained. The ϕ estimate is calculated just as for concurrent validity.

Part 3

Field Survey Methodology*

A variety of Army installations were visited in order to survey the application of criterion-referenced testing techniques in the military. Information was collected to supplement the literature search and review, to provide detailed material on CRT development and use in the Army, and to obtain information on attitudes and opinions of Army testing personnel.

Specifically, the survey gathered data on:

1. How CRTs are developed for Army applications.

2. How CRTs are administered in various Army contexts.

3. How CRT results are used in the Army.

4. Extent of criterion-referenced testing in the Army.

5. The level of personnel who will use the CRT Construction Manual developed during the present project.

6. Problems encountered by Army testing personnel in the development and use of CRTs.

7. Attitudes of Army testing personnel toward the development and use of CRTs.

8. Opinions on the probable future course of criterion-referenced testing in the Army.

9. Sample Army CRTs and problems in developing and using them.

An interview protocol was developed for on-site use at Army posts, to enable standardized collection of information pertaining to the topics listed above. Development of the protocol included several review phases during which revised versions of the protocol were prepared. The final protocol combined separate versions for test constructors, test users, and supervisory personnel; and included several optional items for use in interviews with personnel who were especially knowledgeable about criterion-referenced testing. Thus, the final protocol had a high degree

---

* This section is a brief summary of the methodology used to survey CRT development and use in the Army. For a more detailed description of the methodology, see Swezey, Pearlstein and Ton, 1974.

of utility, and was flexible with respect to the range of topics addressed. Using this protocol, interviews were easily tailored to the ranges of responsibilities, experience, and knowledge possessed by individual interviewees. Appendix A to this report, is a copy of the final version of the protocol.

The interview protocol was used by ASA teams in a series of one-on-one interviews conducted during the months of January, February, and March, 1974. Installations surveyed during this period included the Infantry School at Fort Benning, the Artillery School at Fort Sill, the Air Defense School at Fort Bliss, the Armor School at Fort Knox, and BCT and AIT units at Fort Ord. In addition, test-related departments were surveyed at each post. A total of 105 individuals were interviewed.

ASA survey teams spent three days on-site at each post surveyed. Interviews ranged in duration from approximately one-half to three hours apiece. An average interview took about one and one-half hours. Interview length was at the interviewer's discretion, based on the utility of information obtained from an interviewee.

Personnel in several Combat Arms Schools, MOS testing areas, Training Extension Course (TEC), and Training Center (BCT and AIT) testing programs were interviewed. Figure 3-1 shows the number and types of individuals interviewed in each of these categories. Each interviewee responded to most of the protocol items.

Figure 3-1: Types of Interviewees in the Field Survey

	School	MOS	Training Center	TEC Program	TOTALS
Test developers/ Administrators	41	3	13	4	61
Supervisory Personnel	26	3	11	4	44
TOTALS	67	6	24	8	105

Responses to protocol items that were easily and meaningfully quantifiable were tallied, and percentages of various types of personnel responding in specified ways were computed. Responses to other items that elicited opinion, anecdotal, and process data were summarized by extracting and comparing verbal descriptions.

Part 4

Field Survey Results and Discussion*

Results

Test Construction

Although details of Army test construction processes vary widely across and within Army posts, some general patterns became apparent during the field survey. These include the following:

- Test personnel (both developers and supervisors) are often also involved in preparing objectives, including evaluation standards.
- Practical constraints in the testing situation are frequently considered during test development.
- Although the majority of test personnel interviewed are involved in the actual creation of test items, only a minority create item pools, i.e., write more items than are required for a single form of the test.
- Item analysis techniques are not generally used to select final items for tests. Statistical item analysis techniques are almost never used.
- Test reliability and validity are almost never assessed in a formal manner, and are rarely considered even informally.

Test Administration

A large proportion of interviewees in the survey were involved in administering tests. This is not surprising since much test development is done by school instructors; thus, individuals who create test items also administer the tests in their classes. It was also found that an "assist" method of scoring is frequently used. Test administrators often find it appropriate to provide help to individuals taking the test. The assist method is often used in cases where the examinee could not otherwise complete the test (e.g., a checkout procedure).

---

* This section is a summary and discussion of the results from the field survey of CRT development and use in the Army. A more detailed compendium of the results is provided in this project's Interim Report (Swezey, Pearlstein and Ton, 1974).

Less than half of the 100 interviews queried said that they used go no-go scoring standards on their tests. This does not imply that more than half of the individuals in our survey necessarily use normative scoring standards; instead many use point scales--some of which are criterion-referenced--for scoring.

Many cases in which retesting is done as a matter of course, were cited. For example, in BCT, AIT, and other hands-on performance testing situations, trainees are often given second and third chances to pass particular performance items. Considerably less than half of the interviewees questioned said that they were familiar with team performance testing situations, but many indicated that team performace testing is often individual evaluation in a team context. The actual testing of team performance on the Army posts visited, is very limited.

Using Test Results

The survey found that the most common uses of test results, other than for evaluation of trainee performance, are for improving training and for diagnostic purposes. Seventy-two percent of the interviewees questioned indicated that they use test results for individual diagnostic purposes. Seventy-three percent of the interviewees questioned indicated that they use feedback from tests to improve courses. The way in which this feedback is used, varies widely. For example, some senior instructors indicated that if many trainees from a particular instructor's class perform poorly on certain parts of a test, they would first evaluate the instructor. If several classes taught by different instructors scored poorly on a section of a test, the senior instructor might review the materials used in that portion of the course. In other situations, the test itself is reviewed using feedback from the students.

Finally, less than two-thirds of the interviewees questioned indicated that test results are used to compare trainees, and that such comparisons are not made frequently. It is fortunate that comparisons of this nature are not made more often since the process of making individual comparisons based upon test results is a norm-referenced application.

Types of Tests

The survey discovered that most tests (about 88%) are either paper-and-pencil knowledge tests or hands-on performance tests, as opposed to simulated performance or other types of tests. According to the interviewees, paper-and-pencil knowledge tests account for nearly 50% of those created and used; however, since many interviewees confused paper-and-pencil knowledge tests with paper-and-pencil performance tests, a more realistic estimate is that approximately 25% of the tests are paper-and-pencil knowledge-type, and approximately 25% are paper-and-pencil performance-type (e.g., trajectory computations).

Survey results indicate that nearly three-quarters of the tests constructed or used are performance tests of one sort or another. These results suggest that performance testing has become widespread in many phases of Army evaluation. The survey also showed that tests measuring specific skill and knowledge requirements, and those used at ends of blocks of instruction, account for about 70% of test construction and use. Mid-cycle tests and end-of-course tests together account for less than one-quarter of the tests. According to the interviewees, tests are well distributed throughout instruction, thereby providing frequent feedback and the possibility for on-going remediation.

Problems in Constructing and Using CRTs

Over two-thirds of the interviewees indicated that increased short-term expense may be a problem in the development and use of CRTs, but that in the long run, criterion-referenced testing is less expensive than is norm-referenced testing.

Many individuals in the survey sample felt that time pressures, and to a lesser degree other constraints, often prevent successful construction and use of tests; however, time pressures and other constraints do not usually interfere with test administration tasks. Usually, tests are administered satisfactorily despite time pressures.

Interviewee Attitudes on Criterion-Referenced Testing

In general, interviewees were in favor of the Army trend toward criterion-referenced testing. Eighty-eight percent of the individuals responding, felt that criterion-referenced testing should receive high or top priority in Army assessment programs. Sixty percent felt that criterion-referenced tests should replace most or all norm-referenced tests.

All interviewees felt that criterion-referenced testing is practical and useful in measuring job performance skills. No other item on the survey protocol elicited a 100% positive response.

Discussion

Although criterion-referenced testing is used in today's Army, many NRTs are in use also. This is not surprising, since criterion-referenced testing is a relatively new concept. It was apparent from the survey, however, that CRT use is increasing. School implementation of criterion-referenced testing is still in the beginning stages. Some departments are making serious attempts to incorporate CRTs, while others are only minimally involved. Many employ criterion-referenced terminology, but do not produce true CRTs. This is especially true in

"soft skill" areas, such as tactics and leadership. Most academic departments within the four combat arms schools surveyed, indicated that many of their tests, especially the written ones, are graded on a curve.

MOS testing continues to be primarily norm-referenced. While the situational multiple-choice items from which MOS tests are composed may have been developed in a criterion-referenced fashion (i.e., based on objectives), the items appear suspiciously similar to conventional knowledge test questions on the surface. The proposed Enlisted Personnel Management System (EPMS), including the substitution of Skill Qualification Tests (SQTs) for the present MOS tests, will presumably rectify this situation.

Consideration of the CRT concept is being applied in Training Extension Course packages. However, further development and field testing of the concept in conjunction with TEC is necessary before implementation of TEC CRTs becomes a reality.

At Fort Ord, California, CRTs are employed both in Basic Combat Training and in Advanced Individual Training. Advanced Individual Training in diverse areas, such as field wiring and food services, appears to be benefiting from the use of CRTs. Preliminary indications are that more soldiers are being evaluated more effectively through the application of criterion-referenced testing.

In general, although criterion-referenced testing is not extensive, there are many instances of serious attempts being directed at CRT development and use at the Army installations visited. Implementation of CRTs at first appeared dramatic. But, many of the personnel interviewed confused CRTs with "hands-on" performance testing. In order to be called criterion-referenced, test items must be matched to objectives which are derived from valid performance data. This is not the case for a significant proportion of the "hands-on" performance tests presently used at the sites surveyed.

On the Army posts surveyed, there was much respect for the utility and practicality of criterion-referenced testing. Despite this high regard, there was too little rigorous development or application of CRTs. While progress is being made toward achieving rigor in "hard skill" areas, especially in equipment-related skills, attempts in "soft skill" areas are lacking. This is understandable, since genuine difficulties in specifying soft skill objectives explicitly are often encountered.

Interviewees at all levels indicated a need for increased development and use of criterion-referenced testing in the Army. Many of those indicated that a simple, practical CRT construction manual would consequently be well received at all levels in test development and evaluation units.

A number of difficulties in CRT development and use were observed and/or described during the survey. First, the development of CRTs must be derived from well specified objectives which are, in turn, the results of careful task-analyses. Unfortunately, task analysis data are not available in many cases, and in cases where they are available, they are often disregarded.

The CRT survey suggested that practical constraints for task objectives are usually assessed informally. Frequently, practical constraints to the testing situation are considered only as an afterthought. Constraints which operate in the testing situation should rightfully be considered while a test is being developed. Some Soldier's Manual Army Testing (SMART) books for example, show a minimal regard for practical testing constraints. They contain lengthy checklists which, although possibly of use in evaluating an individual's performance, cannot be followed by test administrators. The problem of failing to consider practical testing constraints adequately may be solved by training test developers to consider such factors as an integral part of the test development process.

Test developers seem to have little difficulty creating items if performances, standards, and conditions are accurately specified in the objectives. However, many Army test developers surveyed indicated that they wrote only the precise number of items required for a specific test. Rarely are extra items written. Items are typically reviewed by subject matter experts and/or test evaluation personnel, and are then revised. Accordingly, there is no empirical selection process for final test items.

Creating a test item pool should become a standard part of the test development process. If twice as many items are developed as are needed for a specific test, the test can be tried out and the final items selected empirically.

A poorly administered test defeats long hours of careful test development. The CRT survey indicated that lack of standardized testing conditions exist in many areas. Careful instructions in test administration are necessary to insure accurate testing. Steps should be taken to insure that test administration practices are clearly defined for each test, and that test administrators are adequately trained.

Finally, a major omission in the development of CRTs, as observed during the Army survey, is the lack of test evaluation. There was virtually no consideration of test reliability and/or validity, although a small subset of interviewees stated that they considered content validity. Army test developers should be instructed in techniques for establishing reliability, and both content and empirical validities of CRTs. Even if a test evidences content validity as a function of careful creation based upon task objectives, reliability is still in question.

Part 5

Developing the CRT Construction Manual

ASA began development of the CRT construction manual by considering both the information on the state-of-the-art gained by the literature review, and the information on Army testing needs, as determined by the field survey. Based on these considerations, a content outline of the manual was prepared. This outline was submitted for review as a part of the project Interim Report (Swezey, Pearlstein, and Ton, 1974). Feedback on the proposed contents was obtained from the COTR and his staff, Army Post Educational Advisors, and other reviewers. The outline was revised according to these inputs.

In order to produce a document presenting "how-to-do-it" procedures for the construction of CRTs, which would be easily understandable by officers and senior enlisted personnel who have little background in psychometrics, the manual was prepared in accordance with the following objectives:

1. Careful structuring to present one point at a time. Each point should involve one, "how-to-do-it" operation.

2. Clear, concise, and straightforward text. Everyday terminology should be used whenever possible, rather than specialized terminology. When psychometric terms were used, they were introduced as needed in an operation. A glossary of psychometric terminology was also included.

3. Practical examples drawn from real life Army situations were used, in lieu of abstract discussions. Theoretical discussions were avoided entirely.

The initial draft of the CRT construction manual required four calendar months to prepare. Following its preparation, it was reviewed by a number of individuals including the COTR and his staff, representatives of the Combat Arms Training Board, representatives of Florida State University's Center for Educational Technology, Dr. Harold Edgerton (consulting for ASA), and a psychometric consultant selected by the COTR.

These reviewers carefully examined the draft manual for both content and structure, and submitted suggested revisions to ASA over a two-month period. ASA collated the suggestions, resolved conflicting suggestions, and, after a thorough in-house editorial review, revised the draft manual. The revised draft manual, entitled Developing Criterion-Referenced Tests (Swezey and Pearlstein, 1974), was printed and distributed for field try-outs and reviews.

Part 6

Field Review Methodology, Results and Discussion

The purpose of this section is to describe the way in which the CRT Construction Manual was evaluated in the field, and the results of the field evaluation. Additionally, this section presents a discussion of the field evaluation findings, in terms of implications for further refinement of the manual.

Methodology

Two versions of a field review package for use in evaluating the CRT Construction Manual were prepared. One version (Form 1) was designed for use by Army test construction personnel, while the other (Form 2) was designed for use by Army educational advisors and by personnel in Army test evaluation units. Both versions included an explanatory cover letter and an evaluation form. Evaluation Form 1 was intended to summarize the utility of the manual, as evaluated by test construction personnel who actually created CRTs using the manual step-by-step. Form 2 was intended to summarize the manual's suitability for the target population, as assessed by Army test and education experts who read the manual in detail. Both evaluation forms consisted of two sections: The first asked for demographic and background data on the respondent, while the second consisted of 35 statements concerning specific aspects of the manual. Respondents were asked to indicate their level of agreement with each statement. Respondents were also requested to include additional comments to elaborate on their evaluations, as necessary. Evaluators using Form 1 packages were asked to send copies of CRTs developed in conjunction with the manual. Appendix B to this report presents copies of both versions of the field review packages.

Field review packages were distributed to the following Army installations:

1. Combat Arms Training Board, Fort Benning, Georgia
2. Infantry School, Fort Benning, Georgia
3. Air Defense School, Fort Bliss, Texas
4. Armor School, Fort Knox, Kentucky
5. Signal School, Fort Gordon, Georgia
6. Basic Combat Training Unit, Fort Ord, California
7. Artillery School, Fort Sill, Oklahoma.

Four copies of the Form 1 package and three copies of the Form 2 package were distributed to each of the above installations. One person at each Army post (chosen on the basis of familiarity with on-post test

construction personnel, test evaluation personnel, and educational advisors) was instructed on distribution of the field review packages at his installation. These persons were also sent cover letters summarizing the distribution procedure. A copy of this cover letter is included as Appendix C to this report.

One copy of the field review package was also sent to Fort Benjamin Harrison, Indiana.

All facilities had approximately one month during which to use, review, and evaluate the CRT Construction Manuals. Follow-up telephone calls were made to respondents whose comments required clarification. In addition, a field visit to Fort Gordon, Georgia was made to observe the field review at the Signal School.

Results and Discussion

Figure 6-1 shows the number of evaluation forms returned. A total of 38 respondents submitted field evaluation forms. Figure 6-2 summarizes the respondents, as to version of the evaluation form used, rank or title, and position. Test construction experience of Form 1 users ranged from 6 months to 25 years, with a mean of 6.5 years. Form 2 users' experience with test construction ranged from 2 years to 35 years, with a mean of 16.2 years.

Figure 6-3 shows Form 1 users' responses (in terms of percentages) to questions 3, 4, and 5. The sample of Form 1 users had high familiarity with CRTs but, many more had developed CRTs than had used those developed by others.

Figure 6-4 shows the percent of types of responses to questions 3, 4a, and 4b of Form 2 users. The responses to question 3 showed a similar pattern to the equivalent question (question 4) on Form 1. Interestingly, questions 4a and 4b indicate that the personnel in the sample using Form 2 were often consulted by people having difficulty with CRTs, and that they feel the CRT manual would have been helpful for these people.

Responses to Items 6 through 40 on Evaluation Form 1, and Items 5 through 39 on Evaluation Form 2, form ordinal scales of measurement. Medians were therefore computed to describe the central tendency of responses to these items (Siegel, 1956). Appendix D to this report shows the tallies of responses to Items 6 through 40 on Form 1, and to Items 5 through 39 on Form 2, as well as the median response for each item. Reaction to the CRT Construction Manual was uniformly favorable. The median responses indicated agreement with favorable statements about the CRT manual. It should also be noted that, in every case, the median response was also the modal response.

Figure 6-1: Field Review Evaluations Forms Returned for Analysis

Facility	Quantity of Evaluation Forms
Fort Knox, Kentucky (Armour School)	8*
Fort Bliss, Texas (Air Defense School)	7
Fort Ord, California (BCT and AIT Units)	6
Fort Gordon, Georgia (Signal School)	5
Fort Sill, Oklahoma (Artillery School)	8*
Fort Benning, Georgia (Infantry School)	4
	Total: 38

*Although only 7 forms were sent to each facility, the Armour School and Artillery School reproduced copies to permit additional, interested test construction personnel to respond.

Figure 6-2: Classification of Field Review Evaluation Respondents

	Form 1	Form 2	Totals
Instructors (including senior instructors)			
Civilian	3		3
Non-Commissioned personnel	5		5
Officers	4	1	5
Education Specialists (Post Educational Advisors, Training Specialists, Education Counselors and MOS Specialists)			
Civilian	5	13	18
Non-Commissioned personnel			
Officers			
Supervisory Personnel (Branch and Division Chiefs and Managers)			
Civilian	1	2	3
Officers	3	1	4
TOTALS	21	17	38

Figure 6-3: Percent* Responses to Q3, Q4, and Q5 on Form 1 (N = 19)

Question	Response		
	Yes	No	No Response
3. Prior to reading the CRT Construction Manual, did you know what a CRT (criterion-referenced test) was?	86%	10.5%	4.5%
4. Have you ever developed a CRT before?	76%	19%	5%
5. Have you ever used a CRT developed by someone else?	52%	43%	5%

*Rounded to nearest half percent.

Figure 6-4: Percent* Responses to Q3, Q4a, and Q4b on Form 2 (N = 15)

Question	Response		
	Yes	No	No Response
3. Have you ever developed (or supervised development of) criterion-referenced tests (CRTs)?	82%	12%	6%
4a. Have you ever been consulted by someone having difficulty in constructing or using a CRT?	82%	12%	6%
4b. If so, do you think the CRT manual would have helped them overcome the problem?	82%	0	18%

*Rounded to nearest half percent.

Respondents using Form 2 agreed strongly with many more items than did Form 1 respondents, indicating that test evaluators and educational experts were even more enthusiastic than test constructors (although both groups were favorably impressed by the manual). Responses to Item 12 on Form 1 indicate that the majority of respondents strongly agree that the manual made the distinction between criterion-referenced and norm-referenced testing clear. On Form 2, responses to 37 percent of the 35 statements had a median of 4, strongly agree. Form 2 respondents especially liked Chapter 1 (Introduction), Chapter 6 (CRT administration and scoring), Chapter 7 (Checking reliability and validity), and the appendices.

A few individuals in both groups disagreed with some items.* Consideration of their comments shed light on their points of disagreement. The problems expressed, can be summarized as follows:

1. The manual does not describe how to derive norm-referenced rankings from CRTs. Some people are required to rank class members based on test results. There is no fool-proof way of ranking students based on CRT results; in fact, the manual discourages this practice. One respondent suggested giving individuals who successfully complete a course, NRTs to determine rankings. This may be a useful suggestion.

2. How to develop soft-skill objectives was not covered in the manual. The manual was not intended to cover development of objectives per se, only assessment of their adequacy. Many people experience difficulty in constructing soft-skill CRTs, primarily because they do not have proper objectives upon which to base the tests.

3. The manual was easy to use, but not easy enough. Individuals making comments of this nature indicated that the manual was easy enough for them to use, but that they thought others might experience difficulty with the level at which the manual was presented.

4. Creating soft-skill CRT items was not covered in sufficient detail. Some respondents felt that a separate chapter on soft-skill item construction might be warranted. However, if soft-skill objectives were more explicit, soft-skill items could be constructed in much the same manner as hard-skill items.

---

* About 3.75% of the responses on Form 2 were unfavorable, and about 9% of the responses on Form 1 were unfavorable.

5. The item analysis procedure (using ϕ) is clear, but is probably impractical for use in the field. Some respondents indicated that there is rarely enough time or try-out sample members available, to perform the recommended item analysis procedure. This may be so, but the procedure recommended is the simplest, empirical item analysis technique practicable. Test constructors, administrators, and supervisory personnel should be educated as to the necessity of empirical item selection procedures, such as the one recommended.

6. Empirical procedures for determination of test-retest reliability, and concurrent and predictive validities, are easy to do, but impractical for field use. This difficulty is essentially the same as the previous one. Army test constructors are not accustomed to checking the reliability and validity of their tests, so the problem of educating them as to the necessity for these types of test evaluation is even more pronounced.

7. The square root tables (Appendix D) do not go up high enough. The tables go from 1 to 1,000. An explanation should be provided at the beginning of Appendix D on how to use the tables to find the square roots of numbers greater than 1,000.

8. The manual is too lengthy. Only a couple of respondents felt the manual is too long. Nevertheless, some consideration should be given to the development of a condensed version of the manual.

9. Technical terminology, though kept at a minimum, may conflict with other terms in use currently. This problem is nearly insoluble, since there are so many terms in use for the same concept throughout the military. The manual does, however, provide a glossary with synonyms.

10. The emphasis on unitary objectives is misleading. It tends to imply an emphasis on testing at low levels of task integration. A related problem is that the emphasis in the manual on responses, rather than on cues (questions) also seems to imply testing at low levels of task integration.

    All levels of task integration were discussed in the manual. What is a unitary objective at a low level of task integration might well be a part of a more complex objective at a higher level of task integration. Similarly, an appropriate cue at a low level of task integration would probably be inappropriate at a higher level of task integration.

11. The manual emphasizes full fidelity testing too much. In addition, it does not stress the importance of "psychological fidelity." This may be so, but, given the lack of explicit rules on when, where, and how to reduce the fidelity of tests, it does not seem appropriate to suggest alternative approaches.

12. The manual is not sufficiently critical of rating scale techniques. There are many reasons why rating scales have not worked well in the past, especially those scales which deal with judgments of global behaviors in work settings. One of the most important of these reasons is that people are unwilling to pass judgments on co-workers or subordinates. Since the manual stresses that, if rating scales are used, they should be behaviorally anchored, and should be referenced to discrete, rather than global behaviors, this difficulty is largely eliminated. Nevertheless, whether or not rating scales can be used effectively with criterion-referenced tests of performance-based training is a matter for additional study.

The vast majority of comments appended to the evaluation forms were favorable. Out of the 17 test construction personnel who responded to Item 40 on Form 1, 14 indicated that they plan to use the procedures presented in the manual when constructing CRTs in the future, and many appended comments reflected this enthusiasm. The following comments are representative:

- "Improvement over usual format for such publications . . . language [is] direct and simple . . . manual is comprehensive."
- "This manual is clear and easy to read . . ."
- ". . . I am an education specialist [and] have been impressed with the overview I have made . . . would like very much to have two [additional] copies of the manual."
- "The manual in its present form could be used as a reference text in a course on test construction conducted at a service school."
- "The manual is extremely comprehensive and does not appear to be lacking any necessary information. It is also very clear and well written. It should be very easy to use."
- ". . . seems to be excellently organized and in clear, precise terms."
- ". . . comprehensive and extremely well written . . . You are to be commended on the fine job . . . will become a very significant and valuable addition to our Army literature on test design."

- ". . . provides the kind of information a 'how-to' manual should provide."

Six individuals sent ASA copies of CRTs, and associated materials, that they created in conjunction with their review of the CRT manual. Four of these CRTs were in hard-skill areas, and two were in soft-skill areas. Test constructors, who used the manual to guide them in creating CRTs, achieved impressive results for the most part. Of special note was a soft-skill CRT and supporting documents that comprised a package of 23 typewritten pages, and was excellent in concept and implementation.

In addition to comments appended to the evaluation forms, ASA received many favorable comments from unsolicited sources in both military and civilian spheres. Nearly 20 such individuals, to whom we did not directly send the manual, contacted ASA to mention their favorable impressions with the manual.

Part 7

Recommendations

The purpose of this section is to present recommendations for future research on, and implementation of criterion-referenced measurement in the Army.

Recommendations for Future CRT Research and Implementation

1. A research effort should be conducted to assess the feasibility of developing and using criterion-referenced MOS tests. Minimally, this research should encompass the following phases:

   A. Outline procedures for converting existing low fidelity, norm-referenced MOS tests to higher fidelity, criterion-referenced MOS tests. These procedures could be based on those presented in the CRT Construction Manual, Developing Criterion-Referenced Tests.

   B. Construct criterion-referenced versions of traditional MOS tests in both hard-skill and soft-skill areas, demonstrating the facility and cost-effectiveness with which such tests can be created using the procedures outlined during phase A above. The criterion-referenced MOS tests should be performance-oriented, at as high a level of fidelity congruent with practicality of administration and maintenance of adequate objectivity. Criterion-referencing of MOS tests should render them more isomorphic to their intended purpose: Assessment of individual performance levels within the occupational specialties.

   C. Compare important psychometric properties of norm-referenced and criterion-referenced tests via field try-out procedures. The comparisons made should include test-retest reliability and both concurrent and predictive validities. In addition, ease of administration and ease of scoring should be compared for the traditional and criterion-referenced versions of the MOS tests. By conducting these various comparisons, cost-benefit analyses of traditional and criterion-referenced MOS tests, in both hard-skill and soft-skill areas, can be computed.

2. Develop more precise objectives for soft-skill areas. A series of research efforts should address the development of operationally-defined objectives, amenable to behavioral assessment, in soft-skill areas. It is inherently more difficult to develop objectives in soft-skill areas than in hard-skill areas. Although the concept of criterion-referenced testing is equally applicable to both areas,

the actual establishment of legitimate objectives is more difficult in areas such as leadership, discipline, tactics, etc., than in more operational areas, such as M16 assembly/disassembly, first aid, etc. Although difficult, the development of soft-skill objectives is certainly possible.

Establishment of performance-oriented objectives in soft-skill areas is frequently time-consuming and tedious. Real ingenuity is also often required. But, although difficult, such objectives can be created.

Research efforts should develop and demonstrate specific techniques for creating adequate soft-skill objectives. These techniques would find an appreciative audience in the Army, as indicated by comments received during conduct of the present study.

3. Implement a program to instruct all Army training and evaluation oriented personnel in Criterion-Referenced Testing. It is apparent that much attention is devoted to CRT concepts in the Army. Yet few individuals are actually familiar enough with these concepts to use them properly. A program should be implemented which will provide training in Criterion-Referenced Testing for persons at all levels in the Army hierarchy who are concerned with the development of procedures for evaluating performance. Special emphasis should be placed upon the necessity for empirical item selection procedures (i.e., item analysis) and empirical evaluation of CRTs' reliability and validity.

4. Implement a program to train test administrators in standardization of testing conditions. Performance test administration is often affected by lack of standardization in testing conditions. This has been particularly true of several tests observed during the Army CRT survey. It is possible, and indeed probable, that failure rates among trainees vary as a function of artifacts in test administration. The CRT manual itself could be used in conjunction with this program.

5. Research should be initiated to investigate the general areas of simulation fidelity in performance testing. It is sometimes the case that high fidelity, high cost, performance simulations are used for testing purposes in areas where lower fidelity simulations may be equally as valid and considerably less expensive. In other areas, low fidelity simulations are used when hands-on testing might be more appropriate.

6. Develop a proceduralized manual to describe techniques appropriate for MOS to SQT (Skill Qualification Test--a criterion-referenced MOS test) conversion and validation. The Enlisted Personnel Management System (EPMS) conference, held at Fort Benjamin Harrison, Indiana on 15-17 October, 1974, resulted in the recommendation that

a manual be developed on how to construct and validate criterion-referenced MOS tests (SQTs). The current manual, Developing Criterion-Referenced Tests, is appropriate, if merged with the recently-generated Item Writer's Guide, and modified to be specifically oriented to MOS tests.

7. Document procedures for developing soft-skill objectives. As noted in the discussion on the Field Review of the CRT Construction Manual, some Army test developers have difficulty in constructing soft-skill CRTs, primarily because they do not have adequate soft-skill objectives from which to work.

8. Develop procedures for using NRTs (or other indices) in conjunction with CRTs. Many Army test developers are concerned about the requirement that they provide norm-referenced information on examinees who have been tested by CRTs. This is a genuine problem with no easy resolution. Procedures for simultaneously using CRT and other, norm-referenced indices should be developed for situations requiring both norm-referenced decisions and criterion-referenced decisions.

9. Develop a condensed version of the CRT Manual. A condensed version of the CRT construction manual would be valuable for personnel who are already fairly familiar with CR testing. This version should omit much of the detail (and introductory material) presented in Developing Criterion-Referenced Tests.

APPENDIX A

Interview Protocol:

Survey of Criterion-Referenced

Testing in the Army

Interviewer Statement: Now, I would like to discuss with you, some tasks that may be involved in test construction and use. These tasks are done in different ways in different places. Sometimes they are combined, in other cases some are eliminated. They often go by different names. Would you please tell me which of these you are involved in.

*4. Writing objectives. That is--determining what the test will measure and the conditions under which the measurement will occur in terms of precise, behavioral statements.

Have you been involved in writing objectives? Yes_____ No_____

If yes, (a) how long have you been doing this? Years_____ Months_____

(b) do you write objectives in operational, behavioral terms?

Yes_____ No_____ Don't understand_____

*5. Setting standards. That is--defining the standards against which performance is evaluated. In many cases, these standards are very similar to the stated objectives.

Have you participated in setting standards? Yes_____ No_____

If yes, how long have you been doing this? Years_____ Months_____

*6. Imposing practical constraints. That is--deciding how the test must be built so it can actually be used within the limits of the situation for which it is designed. For example, there are often time constraints involved in testing complex skills.

Have you been involved in this? Yes_____ No_____

If yes, how long have you been doing this? Years_____ Months_____

* 11. Measuring reliability. That is--determining if a test will give similar scores when measuring similar performance. For example, a person taking equivalent versions of the same test should score about the same on both, if he has had no practice in between.

Have you been involved in measuring the reliability of tests? Yes____ No____

If yes, (a) how long have you been involved in measuring reliability?

Years_____ Months_____

(b) do you compute coefficients of reliability?

Yes_____ No_____ Don't know_____

* 12. Evaluating validity. The test developer must determine whether the test is actually measuring what it is supposed to measure. Personnel who score high on the test should also perform very well on the task that test is supposed to measure, while those who score low should not be able to perform the task as well.

Have you helped to validate tests? Yes_____ No_____

If yes, (a) how long have you been doing so? Years_____ Months_____

(b) do you use content validity as opposed to predictive validity?

Yes_____ No_____ Don't know_____

13. Scoring. How are tests generally scored? Are norms set as standards using bell shaped curves, or are "go-no go" type standards used?

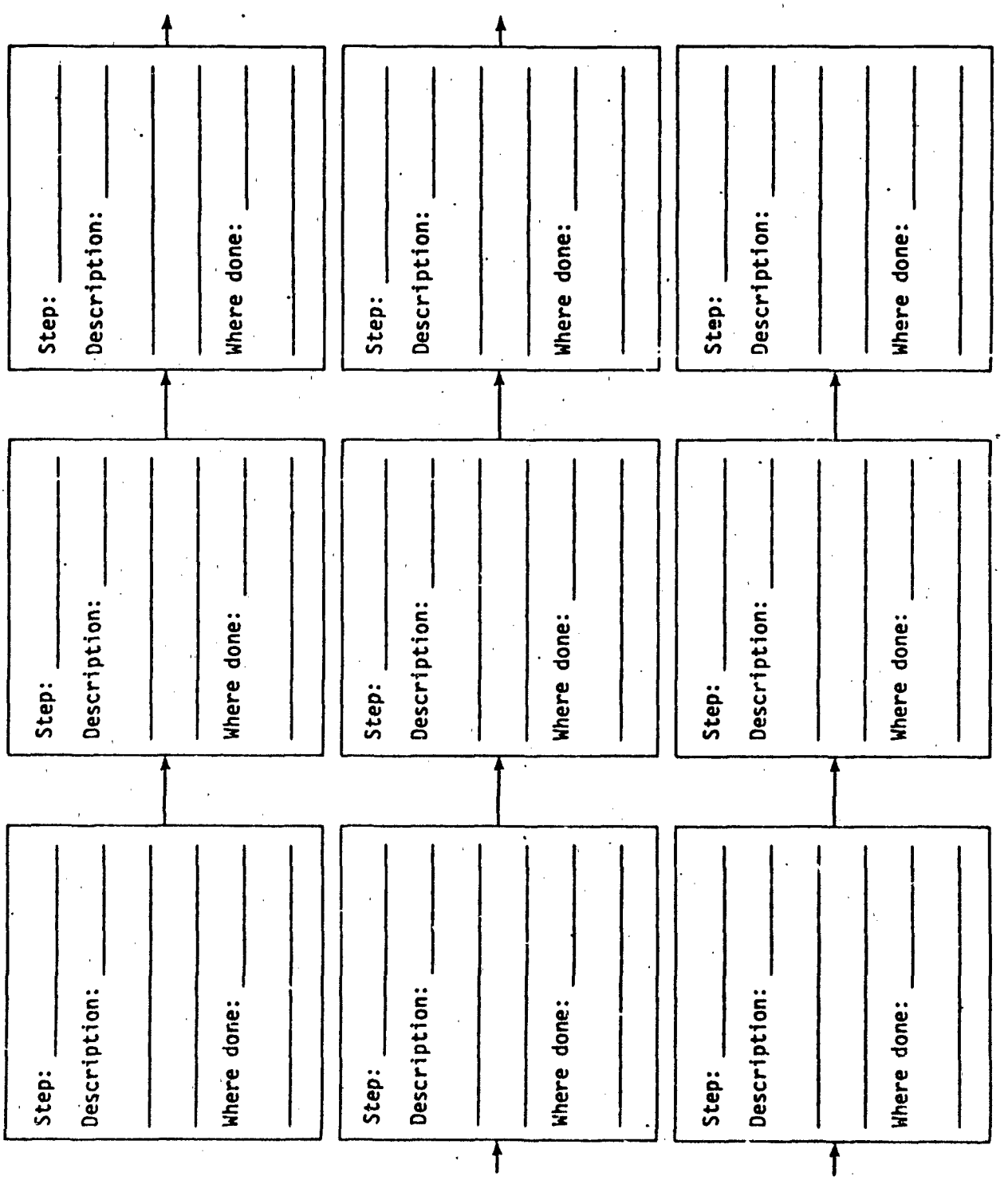
Norms_____ go-no go_____ Other_____

_____

_____

_____

*18. Let's consider the overall test development and use process. Would you help me fill in the steps, as they actually happen at this post in developing and using tests? Since you may not participate in all steps yourself, we'd like to determine who does what step where.



Interviewer Statement: Now I would like to discuss some of the tasks that you're involved in.

19. What inputs do you have available in terms of documents, data, job aids, field manuals, etc.? REQUEST THESE

20. Which of these inputs do you actually use?

*21. [If answer to 20 is other than "all of them", interviewer asks #21] Why do you use these and not the others?

22. What products do you prepare? REQUEST THESE

23. How are these outputs used?

24. What problems have you encountered?

25. How did you resolve these problems?

_____

_____

_____

*26. Is any special training available for testing personnel? Yes_____ No_____

If yes, please briefly describe this training?

_____

_____

_____

27. What proportion of the tests you have participated in making or using are:

A. Paper-and-pencil knowledge tests? _____

B. Simulated performance tests? E.g., using mockups and drawings _____

C. "Hands on" performance tests? _____

D. Other? Specify: _____ _____

_____

What proportion of the tests you have participated in making or using are for:

A. Specific skill and knowledge requirements? _____

B. Specialty areas in a course? _____

C. End of block within a course? _____

D. Mid cycle within a course? _____

E. End of course? _____

*28. Are you familiar with any team performance situations that were evaluated by tests? Yes_____ No_____

*29. Would you briefly describe how tests were used to measure team performance?

_____

_____

30. Have time pressures, or other constraints, prevented you from successfully carrying out some of the tasks involved in test construction and use?

Yes_____ No_____

If yes, describe how you were affected by a constraint.

_____

_____

_____

*31. Can you describe any cases in which tests were developed which were not suitable, in your opinion, for the intended uses? Yes_____ No_____

Description: _____

_____

_____

If it is the interviewer's opinion that interviewee does not understand the distinction between Criterion-Referenced Testing and norm-referenced testing:

STOP HERE

Otherwise go on.

32. One of the main purposes of our work for the Army is to develop a manual on how to construct Criterion-Referenced as opposed to Norm-Referenced Tests. Who will be the primary users of a manual of this type on this post?

_____

_____

_____

*33. As you know, in recent years the Army has put increasing emphasis on using Criterion-Referenced Tests in appropriate testing situations. There is still much disagreement, though, about what a Criterion-Referenced Test really is. How is the term "Criterion-Referenced Test" used on this post?

_____

_____

_____

*34. How strongly do you feel about future use of Criterion-Referenced Testing in the Army? Should Criterion-Referenced Test development receive high or low priority in terms of Army assessment programs?

_____ Strongly against--Criterion-Referenced Testing should receive bottom priority, or dropped entirely.

_____ Against--Criterion-Referenced Testing should receive low priority.

_____ Neutral--Criterion-Referenced Testing should receive average priority.

_____ For--Criterion-Referenced Testing should receive high priority.

_____ Strongly for--Criterion-Referenced Testing should receive top priority, Criterion-Referenced Tests should replace most or all norm-referenced tests.

*35. Do you think cost is a major factor in determining whether Criterion-Referenced Tests are developed and administered in the Army? That is--have you found that Criterion-Referenced Tests are more or less expensive to develop and administer than conventional, norm-referenced tests?

Less expensive_____ About the same_____ More expensive_____

*36. Could you describe a situation in which a Criterion-Referenced Test was found to be prohibitively expensive to develop?

_____

_____

_____

37. Do you think that there are any particular advantages or disadvantages to developing and using Criterion-Referenced tests in the Army (as opposed to norm-referenced measures)? Yes_____ No_____

What are some advantages or disadvantages?

_____

_____

_____

38. Are there any special problems you have encountered while developing or using Criterion-Referenced Tests, as opposed to problems normally encountered with norm-referenced tests? Yes_____ No_____

If yes, describe these special problems and how you overcome them:

_____

_____

_____

*39. How serious are these problems? That is, how much do they affect the overall accomplishment of testing objectives?

_____

_____

40. Do you feel that Criterion-Referenced Testing is practical and useful in measuring job performance skills? Yes_____ No_____

Why? _____

_____

_____

*41. Are there other areas (such as knowledge tests and achievement tests) where this concept could be useful? Yes_____ No_____

Why? _____

_____

_____

42. What should we include to make the manual useful?

_____

_____

_____

APPENDIX B

Field Review Evaluation Packages:

Form 1 and Form 2

Form 1

DRAFT LETTER
(for use with Evaluation Form #1)

Dear Sir,

Applied Science Associates, Inc. (ASA) wishes to solicit your aid in assessing the enclosed version of a criterion-referenced test construction manual entitled Developing Criterion-Referenced Tests. This manual, developed under contract No. DAHC19-74-C-0018 for the Army Research Institute for the Behavioral and Social Sciences is intended to aid Army test developers in the construction of criterion-referenced tests (CRTs). Your comments and suggestions will be used to help revise this version of the manual.

Here is how you can help:

1. Read the manual, familiarizing yourself with its contents.

2. Develop a CRT of your own (for whatever use is appropriate to your testing needs), following the procedures presented in the manual. Use the manual step-by-step as you develop this test.

3. Fill out the enclosed evaluation form indicating how useful the manual was in guiding you through the test construction process. Feel free to include additional comments which you think would be helpful to us for revising the manual.

4. Send a copy of the test you constructed, and associated documentation if possible, along with the completed evaluation form to:

APPLIED SCIENCE ASSOCIATES, INC.
11800 Sunrise Valley Drive
Reston, VA 22091

If ASA receives your materials by 1 November, 1974, we will be able to consider your evaluation within the time constraints imposed by the contract.

Please call ASA at (703) 620-3494 if you have any questions.

Evaluation Form 1

Please use this evaluation form to indicate how helpful the CRT Manual was in guiding you through the test construction process.

Name: _____ _____ _____ _____
Rank or Title | First | Middle Initial | Last

Address on Post: _____ _____
Bldg & Number | Street Address, if applicable

_____ _____ _____ _____
Post | City | State | Zip Code

Phone Number: _____ _____
Area Code | Number on Post at which you can be reached

1. What is your position? [for example: Senior Instructor, Nuclear-biological-chemical committee]

   _____

2. How long have you been involved with some aspect of test construction?

   _____ _____
   years | months

3. Prior to reading the CRT construction manual, did you know what a CRT (criterion-referenced test) was?

   Yes No Circle one.

4. Have you ever developed a CRT before? Yes No Circle one.

5. Have you ever used a CRT developed by someone else?

   Yes No Circle one.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 3

1	2	3	4	18.	The concept of practical constraints, and how they may constrain testing of all objectives as stated, was adequately presented.
1	2	3	4	19.	How to overcome practical constraints--either by selecting among objectives or by modifying objectives in light of the constraints--was clear.
1	2	3	4	20.	When and how to sample items for objectives was easily understandable.
1	2	3	4	21.	Testing under multiple conditions and how to sample multiple conditions was clear.
1	2	3	4	22.	The guidelines for determining how many items to include in a CRT were helpful.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 4

1	2	3	4	23.	The explanation of how to create items based on test plan specifications was adequate.
1	2	3	4	24.	The material concerning developing specific and general test instructions was helpful and at the right level of detail.
1	2	3	4	25.	The section on assessing adequacy of items was clear and useful.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 5

1	2	3	4	26.	How to select a proper try-out sample and conduct an item pool try-out was clear and easy to follow.
1	2	3	4	27.	Computing item analysis values using ϕ on try-out results was presented in a clear fashion.
1	2	3	4	28.	How to reduce the item pool by considering try-out results, item analysis, and item reviews was presented adequately.
1	2	3	4	29.	What to do if too few or too many items were left after item analysis and reviews was clear.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 6

1	2	3	4	30.	The material on standardizing test administration procedures and administering CRTs was clear and useful.
1	2	3	4	31.	The information on how to score CRTs, establish cut-off scores, and report test results was adequate.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 7

1 2 3 4	32. The procedure for determining test-retest reliability was clear and easy to follow.
1 2 3 4	33. How to assess content validity was clear.
1 2 3 4	34. How to determine concurrent validity was presented adequately.
1 2 3 4	35. How to determine predictive validity was presented adequately.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN APPENDICES

1 2 3 4	36. Appendix A (Checklist for constructing CRTs) was useful.
1 2 3 4	37. Appendix B (Checklist for evaluating CRTs) would be helpful in evaluating CRTs that have already been developed.
1 2 3 4	38. Appendix C (Glossary) was helpful and contained all terms I needed to look up.
1 2 3 4	39. Appendix D (Square root tables) was useful in calculating values of ϕ .
1 2 3 4	40. I plan to use the procedures presented in this manual when developing CRTs in the future.

PLEASE FEEL FREE TO INCLUDE ADDITIONAL COMMENTS (ATTACH SEPARATE SHEETS AS NECESSARY).

Form 2

DRAFT LETTER
(for use with Evaluation Form #2)

Dear Sir,

Applied Science Associates, Inc. (ASA) wishes to solicit your aid in assessing the enclosed version of a criterion-referenced test construction manual entitled Developing Criterion-Referenced Tests. This manual, developed under contract No. DAHC19-74-C-0018 for the Army Research Institute for the Behavioral and Social Sciences is intended to aid Army test developers in the construction of criterion-referenced tests (CRTs). Its target audience is composed of senior enlisted personnel and officers who are involved in test construction, but who may not be sophisticated with respect to psychometric techniques.

Your comments and suggestions will be used to help revise this version of the manual.

Here is how you can help:

1. Read the manual.

2. Complete the enclosed evaluation form to evaluate the suitability of the manual.

3. Feel fre to include any additional comments which you think would be helpful to us for revising the manual.

Please send the completed evaluation form and any additional materials to:

APPLIED SCIENCE ASSOCIATES, INC.
11800 Sunrise Valley Drive
Reston, VA 22091

In order to be able to use your evaluation within the time constraints imposed by the contract, ASA must receive your inputs by 1 November, 1974.

Please call ASA at (703) 620-3494 if you have any questions.

Evaluation Form 2

Please use this evaluation form to indicate how useful you think the CRT Manual will be for Army Test Constructors.

Name: _____ _____ _____ _____
Rank or Title | First | Middle Initial | Last

Address on Post: _____ _____
Bldg & Number | Street Address, if applicable

_____ _____ _____ _____
Post | City | State | Zip Code

Phone Number: _____ _____
Area Code | Number on Post at which you can be reached

1. What is your position? [for example: Post Educational Advisor]

   _____

2. How long have you been involved with test construction and related issues?

   _____ _____
   years | months

3. Have you ever developed (or supervised development of) criterion-referenced tests (CRTs)?

   Yes No Circle one.

4. Have you ever been consulted by someone having difficulty in constructing or using a CRT?

   Yes No Circle one.

   If so, do you think the CRT manual would have helped them overcome the problem?

   Yes No Circle one.

Directions: The remainder of this evaluation form consists of statements about the manual, Developing Criterion-Referenced Tests. Each statement is preceded by the numbers 1 through 4.

Circle 1 if you strongly disagree with the statement.
Circle 2 if you disagree with the statement.
Circle 3 if you agree with the statement.
Circle 4 if you strongly agree with the statement.

Please respond to each statement, circling the number which best expresses your opinion. Remember the manual's audience may be composed of people who are not sophisticated with respect to psychometric concepts and terminology.

1 2 3 4	5. The manual would be very helpful in guiding people through the CRT construction process.
1 2 3 4	6. The manual would be easy for Army test developers to use.
1 2 3 4	7. Examples provided in the manual are useful.
1 2 3 4	8. The manual covers all the points it should.
1 2 3 4	9. I would recommend that this manual be used by Army test developers whenever possible.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 1

1 2 3 4	10. The concept of criterion-referenced testing is explained clearly.
1 2 3 4	11. The explanation of when to develop CRTs is clear and accurate.
1 2 3 4	12. The distinctions between criterion-referenced and norm-referenced testing are clear.
1 2 3 4	13. The overview of the CRT construction process provides a clear idea of what the manual covers.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 2

1 2 3 4	14. The discussion of the three main parts of an objective is clear and comprehensive.
1 2 3 4	15. The process of establishing unitary objectives is clear.
1 2 3 4	16. The distinctions among overt main intents, covert main intents, and indicators are clear.
1 2 3 4	17. The sequence of operations for assessing the adequacy of objectives is clear and to the point.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 3

1 2 3 4	18. The concept of practical constraints, and how they may constrain testing of all objectives as stated, is adequately presented.
1 2 3 4	19. The procedures for overcoming practical constraints--either by selecting among objectives or by modifying objectives in light of the constraints--are appropriate and presented clearly.
1 2 3 4	20. When and how to sample items for objectives is presented adequately.
1 2 3 4	21. The information on how to test under multiple conditions (including how to sample multiple conditions) is appropriate and clear.
1 2 3 4	22. The guidelines for determining how many items to include in a CRT are helpful.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 4

1 2 3 4	23. How to create items based on test plan specifications is explained adequately.
1 2 3 4	24. The material on developing specific and general test instructions is helpful and at the right level of detail.
1 2 3 4	25. The section on assessing the adequacy of items is clear and useful.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 5

1 2 3 4	26. The procedure for selecting a proper try-out sample and conducting an item pool try-out is clear and easy to follow.
1 2 3 4	27. The presentation of how to do an item analysis using ϕ is clear and appropriate.
1 2 3 4	28. The material on reducing the item pool by considering try-out feedback, item analysis, and item reviews is adequate.
1 2 3 4	29. What to do if too few or too many items remain after reduction of the item pool is clear and appropriate.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 6

1 2 3 4	30. The material on standardizing test administration procedures and administering CRTs is clear and useful.
1 2 3 4	31. How to score CRTs, establish cut-off scores, and report test results are explained adequately.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN CHAPTER 7

1	2	3	4	32.	The procedure for determining test-retest reliability is appropriate and presented clearly.
1	2	3	4	33.	How to assess content validity is clear and practical.
1	2	3	4	34.	How to determine concurrent validity is appropriate and presented clearly.
1	2	3	4	35.	How to determine predictive validity is appropriate and presented clearly.

THE FOLLOWING STATEMENTS CONCERN MATERIAL IN APPENDICES

1	2	3	4	36.	Appendix A (Checklist for constructing CRTs) is useful.
1	2	3	4	37.	Appendix B (Checklist for evaluating CRTs) is useful.
1	2	3	4	38.	Appendix C (Glossary) is useful and covers all necessary terms.
1	2	3	4	39.	Appendix D (Square root tables) is useful and appropriate for this manual.

PLEASE FEEL FREE TO INCLUDE ADDITIONAL COMMENTS (ATTACH SEPARATE SHEETS AS NECESSARY)

APPENDIX C

Cover Letter to Contact Man at Each Post Describing How Materials Are To Be Distributed

Dear _____,

In accordance with our recent telephone conversation, enclosed are the materials you need to help in evaluating the CRT construction manual, Developing Criterion-Referenced Tests. Seven (7) copies of the manual, and seven (7) field review packages--each consisting of an explanatory cover letter and an evaluation form, are included.

Please note that four (4) field review packages are labeled "Form 1", and three (3) are labeled "Form 2". Please distribute the packages as follows:

1. Keep one copy of the manual and one "Form 2" field review package for yourself. Follow the directions in the cover letter enclosed in the field review package.

2. Select two (2) people on your post who are experienced in test construction methodology, educational technology, or test evaluation. Give each a copy of the manual and a "Form 2" field review package. Ask them to follow the directions in the cover letter.

3. Select four (4) people on your post who are actively involved in test construction tasks. These may be instructors, senior instructors, etc. Give each a copy of the manual and a "Form 1" field review package. Ask them to follow the directions in the cover letter.

It is important to remember that all completed evaluations must be received by Applied Science Associates, Inc. (ASA) by 1 November, 1974. Consequently, the interval in which evaluations must be completed is relatively brief. To ensure meeting deadlines, it is important that you distribute these materials as soon as possible.

If you have questions concerning appropriate candidates to receive the field review packages and manuals, please feel free to contact ASA at (703) 620-3494. Thank you very much for your cooperation.

Sincerely,

Robert W. Swezey, PhD
Applied Science Associates, Inc.

Richard B. Pearlstein, PhD
Applied Science Associates, Inc.

Angelo Mirabella, PhD
Army Research Institute

C-1

APPENDIX D

Results of Field Review Evaluation:

Tallies of Responses on Form 1

and Form 2

Form 1:

Items 6 - 40

Results of Field Evaluation of CRT Construction Manual

FORM 1

Item No.*	No Response	Response				Median
		1 (Strongly Disagree)	2 (Disagree)	3 (Agree)	4 (Strongly Agree)	
6	III		I	𝍸 𝍸 II	𝍸	3
7	II		IIII	𝍸 𝍸 IIII	I	3
8	II		I	𝍸 𝍸 𝍸 II	I	3
9	I		IIII	𝍸 𝍸 III	III	3
10			I	𝍸 𝍸	𝍸 𝍸	3
11			II	𝍸 𝍸 I	𝍸 III	3
12			I	𝍸 II	𝍸 𝍸 III	4
13				𝍸 𝍸 𝍸	𝍸 I	3
14			I	𝍸 𝍸 I	𝍸 IIII	3
15		I	I	𝍸 𝍸 II	𝍸 II	3
16			I	𝍸 𝍸 𝍸	𝍸	3
17			I	𝍸 𝍸 𝍸	𝍸	3
18				𝍸 𝍸 II	𝍸 IIII	3
19			I	𝍸 𝍸 𝍸 I	IIII	3
20			II	𝍸 𝍸 𝍸 II	II	3
21			III	𝍸 𝍸 III	𝍸	3
22			I	𝍸 𝍸 𝍸 III	II	3
23			II	𝍸 𝍸 𝍸 I	III	3
24	I		II	𝍸 𝍸 III	𝍸	3
25	I	I	II	𝍸 𝍸 𝍸	II	3
26		I	I	𝍸 𝍸 II	𝍸 II	3
27		II	II	𝍸 𝍸 I	𝍸 I	3
28		I	II	𝍸 𝍸 IIII	IIII	3
29		I	I	𝍸 𝍸 𝍸 I	III	3
30			II	𝍸 𝍸 III	𝍸 I	3
31		I	II	𝍸 𝍸 III	𝍸	3
32		I	I	𝍸 𝍸 𝍸 I	III	3
33			II	𝍸 𝍸 IIII	𝍸	3
34			III	𝍸 𝍸 IIII	IIII	3
35	I		II	𝍸 𝍸 III	IIII	3
36			I	𝍸 𝍸 𝍸 II	III	3
37		I		𝍸 𝍸 𝍸 II	III	3
38			I	𝍸 𝍸 𝍸	𝍸	3
39	I	I	II	𝍸 𝍸 II	𝍸	3
40	II	I	II	𝍸 𝍸 III	III 0-2	3

Form 2

Items 5 - 39

Results of Field Evaluation of CRT Construction Manual

FORM 2

Item No.*	No Response	Response				Median
		1 (Strongly Disagree)	2 (Disagree)	3 (Agree)	4 (Strongly Agree)	
5	1			11	11	3
6			4	11	1	3
7			1	11	4	3
8		2	2	10	3	3
9			2	10	5	3
10			1	6	10	4
11				11	6	3
12				10	8	3
13				6	11	4
14			1	3	13	4
15			2	8	7	3
16			1	7	8	3
17				11	4	3
18				5	12	4
19				9	8	3
20				10	7	3
21				10	7	3
22				10	8	3
23				11	4	3
24				9	8	3
25			1	8	8	3
26				11	4	3
27			1	10	9	4
28				10	7	3
29				9	8	3
30	1			8	11	4
31			3	3	10	4
32	1			9	7	3
33	1			6	10	4
34	1			6	10	4
35	1			8	7	3
36				6	9	4
37				6	9	4
38				6	10	4

References

Edgerton, H. A. Personal communication, 1974.

Graham, D. L. An examination of the feasibility of using criterion-referenced measurement in large-scale, survey testing situations. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1974.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement. 1974, 11(2), 93-99.

Handbook for designers of instructional systems. AF Pamphlet 50-58, Wright-Patterson Air Force Base, Ohio, 1973.

Livingston, S. A classical test-theory approach to criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Mager, R. F. Preparing instructional objectives. San Francisco: Fearon, 1962.

Mager, R. F. Measuring instructional intent. San Francisco: Fearon, 1973.

Oakland, T. An evaluation of available models for estimating the reliability and validity of criterion-referenced measures. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Siegel, Sidney. Nonparametric statistics for the behavioral sciences. H. F. Harlow (Ed.) McGraw-Hill Series in Psychology, 1956.

Smith, R. G., Jr. Controlling the Quality of Training. HumRRO Technical Report 65-6, June 1965.

Soldiers' Manual Army Testing (SMART). TRADOC Pamphlet No. 600-9, Fort Monroe, Virginia, 1973.

Swezey, R. W. and Pearlstein, R. B. Developing criterion-referenced tests. Applied Science Associates, Reston, Va., Technical Report 287-AR18(2)-IR-0974-RWS, 1974.

Swezey, R. W., Pearlstein, R. B., and Ton, W. H. Criterion-referenced testing: A discussion of theory and of practice in the army. Applied Science Associates, Reston, Va., Technical Report 273-AR18(1)-IR-0474-RWSW, 1974.

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement. 1974, 11(1), 63-64